Exploring an Explanatory Child Speech Intelligibility Model Using Phonetically Contrasted Word Productions

by

Katherine Eilene Willoughby

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Speech-Language Pathology, Master of Science

Auburn, Alabama
May 5, 2019

Keywords: intelligibility, speech sound disorders,
direct magnitude estimation, crowdsourcing

Approved by

Marisha Speights Atkins, Chair, Assistant Professor of Communication Disorders
Dallin J. Bailey, Assistant Professor of Communication Disorders
Mary J. Sandage, Associate Professor of Communication Disorders
Aurora J. Weaver, Assistant Professor of Communication Disorders

Abstract

Purpose: In clinical practice, there is no standard measure of intelligibility that explains the kinds of difficulties that listeners who are inexperienced with child speech may encounter. This study aimed to investigate phonetic-based error types that occur in the speech of young children that contribute to decreased intelligibility.

Method: Speech recordings of 9 preschool children producing phonetic contrasts that reflect common phonological disorders were analyzed by inexperienced listeners. To investigate the type of difficulties listeners in the general population might encounter, participants were recruited through the crowdsourcing platform Amazon Mechanical Turk (AMT). Testing the effect of phonetically contrasted words on the listeners' ability to recognize the word and rate the intelligibility of the word was housed through a web-based platform compatible with AMT, Intelli-turk©. It was hypothesized that listeners inexperienced with child speech could rate the speech of children using Direct Magnitude Estimation and reflect different levels of intelligibility in agreement with previous word production accuracy measures.

Results: The results of this study support the correlation between measures of whole-word accuracy and ratings of intelligibility. It was also found that different types of errors may contribute to listeners' intelligibility ratings, meaning that listeners inexperienced with child speech productions identify differences in intelligibility categorically. Specific types of errors that contribute to the confusion of listeners' intelligibility rating differed according to speaker

accuracy level and phonetic contrast categories. This preliminary investigation yielded promising results towards establishing an explanatory model of intelligibility for preschool age children.

Acknowledgments

Primarily, I would like to express my gratitude to my committee chair Dr. Marisha Speights Atkins for her leadership, positivity, and consistent support throughout this project. My progress as a researcher, writer, and professional reflects your mentorship. Second-most I would like to thank my thesis committee members: Dr. Dallin Bailey, Dr. Mary Sandage, and Dr. Aurora Weaver for your time, expertise, and insight into my thesis project. I would like to thank Javier Livio and Bill Tran from the Computer Software Engineering Department for their advanced technological contributions without which my project would not have seen success. A special thank you is owed to the members of the Technologies for Speech-Language Research Lab for their hard work and support in completing this project. I would also like to thank my parents, Ian and Liz Willoughby, for instilling in me the principles of hard work, fostering my desire to learn, and for their unyielding encouragement.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

SSD	Speech-Sound Disorder
WRS	Word Recognition Score
DME	Direct Magnitude Estimation
SLP	Speech Language Pathologist
HAPP	Hodson Assessment of Phonological Patterns
PCC	Percentage Consonants Correct
AMT	Amazon Mechanical Turk
WRE	Word Recognition Error
TE	Target Error
PCE	Pair Contrast Error
Non-TE	Non-Target Error
SG	Speaker Group

I. Introduction

Articulation and phonological disorders are among the most widely treated disorders by school speech language pathologists. It is reported that on average 8 to 9 percent of young children have articulation or phonological disorders (NIDCD, 2016). Children with phonological disorders are often reported as unintelligible to inexperienced listeners (Edition & Bauman-Waengler, 2012). Improvement of speech intelligibility is a primary aim in remediation of articulation or phonological disorders and is described as "one of the most fundamental aspects necessary for successful oral communication" (Connolly, 1986 p.371).

Procedures for objective measurement of intelligibility lack agreement. Two common approaches are signal dependent listening tasks and perceptual-based standardized assessments. Signal-dependent intelligibility assessment methods, such as word recognition, direct magnitude estimation, and interval scales, are typically performed by having two to five listeners rate recorded speech samples. Due to time, and availability for listeners these methods are not readily used as regular outcome measures of speech evaluation in clinical settings (Gordon-Brannan & Hodson, 2000; Miller, 2013). Perceptual-based measures, more routinely used in clinical settings rely on clinicians' ability to accurately listen to the child's speech production and analyze the production to determine the severity and implied deficits of intelligibility. However, these are not direct measures of intelligibility and they lack a more detailed analysis of specific characteristics of speech that contribute to decreased intelligibility.

There is no widely accepted measure of intelligibility used to describe degree of impairment in children with speech sound disorders (Hustad, 2018; Miller, 2013). Standardized

assessments for intelligibility include but are not limited to the Assessment of Intelligibilty in Dysarthric Speech (AIDS) and a recent addition to the Goldman-Fristoe third edition (GFTA-3). AIDS tests single words, sentences, and conversation speaking rates of speakers with dysarthira (Yorkston, Beuklemon, & Traynor, 1984). There is no equivalent standardized test for children with speech sound disorders. The GFTA-3 offers a four-point scale rating of intelligibility by the clinician during the sounds-in-sentences test and compares scores to age-expected comprehensive intelligibility percentages (Goldman & Fristoe, 2015). Scaling measures of intelligibility performed by experienced listeners have limitations that should invoke caution to their use in clinical assessment. One such limitation is that speech can often be produced with certain error types and be understood by a listener particularly, those with experience listening to disordered speech (Kent, 1996). Additionally, impressionistic intelligibility scores made during accuracy assessments result in general percentages without explanation of errors causing unintelligibility.

Researchers have explored other more systematic approaches to characterize specific attributes of speech sounds that could explain decreased intelligibility in speakers. Kent (1989) proposed an explanatory speech intelligibility assessment method, the Diagnostic Intelligibility Test, that identified specific phonetic attributes including both consonants and vowels that contribute to intelligibility in adults with dysarthria. This intelligibility test used single word stimuli in phonetic contrast categories to obtain error profiles comparing intelligibility level across acoustic parameters (Weismer, Martin, & Kent, 1992). It was revealed that error profiles differed across speakers regardless of similarity in overall intelligibility scores. These differences suggest that there may be factors not reflected in general percentage-based scores contributing to intelligibility. Although this work has explored contributing factors of intelligibility in adult

speakers, few studies have investigated explanatory models of intelligibility in children with speech sound disorders. The primary aim of this study was to employ word sets from the Diagnostic Intelligibility Test (Kent, et al., 1989) in order to identify the phonetic contrasts that predict decreased speech intelligibility in children with and without speech sound disorders. A secondary aim was to explore the feasibility of crowdsourcing for recruiting inexperienced listeners to provide ratings of intelligibility. In this way, we hope to establish an ecologically sound research approach which aids in the investigation of explanatory intelligibility assessment through the general population of inexperienced listeners.

II. Literature Review

Intelligibility as Related to Speech Sound Disorders

Children with articulation and phonological disorders are often reported as unintelligible to inexperienced listeners (Edition & Bauman-Waengler, 2012). Improvement of speech intelligibility is a primary aim in remediation of articulation or phonological disorders and is described as "*one of the most fundamental aspects necessary for successful oral communication*" (Connolly, 1986 p. 371). Speech intelligibility is essential for functional communication and participation in multiple social environments. Intelligibility measurements have been considered to be a central component of clinical decision making and assessment of the efficacy of treatment (Miller, 2013 p. 601; Hustad, Oakes, & Allison, 2015). Although there is an agreement that measuring intelligibility is crucial in evaluating functional speech, 75% of SLPs estimate intelligibility without the use of any standardized protocol (Skahan, Watson, & Lof, 2007). Clinicians often use adopted percentages of expected intelligibility by age: 3 years 75%, 4 years 85%, and 5 years 95% (Bankson, Bernthal, & Flipsen, 2013) as a method of classifying the degree of intelligibility. These judgments are impressionistic and may result in variability in classification across raters thus being a less reliable means of determining intelligibility in speakers (Kent, 1996). More systematic methods for measurement of intelligibility have been of interest to researchers for several decades however, a lack of consensus on an intelligibility assessment approach remains (Kwiatkowski & Shriberg, 1992; Lousada, Jesus, Hall, & Joffe, 2014; McLeod, Harrison, & McCormack, 2012 ; Speake et al,. 2012).

Current Practices in Intelligibility Measurement

A national survey of school-based speech-language pathologists revealed that the Hodsen Assessment of Phonological Pattern (HAPP) and Percent Consonant Correct (PCC), were the most commonly used intelligibility assessment tools (Logan, 2010). These assessments have traditionally been considered severity measurements which yield information reflecting accuracy and frequency of errors. The gold standard for speech intelligibility assessment, as reported by the American Speech and Hearing Association ("ASHA Practice Portal: Speech Sound Disorders-Articulation and Phonology", n.d.), is perceptually based judgments that aid in determining the severity of a speech sound disorder but focus on the listener's ability to decode the acoustic signal as the speaker's intended word. To measure intelligibility, listeners are presented with pre-recorded speech samples and asked to identify what they heard. Intelligibility judgments are made based on the listeners' perception of the speech sample. Responses, typically of 3-5 listeners, are scored based on the number of words matched correctly and yield a percent-intelligibility score (Gordon-Brannan & Hodson, 2000; Kent, Miolo, & Bloedel, 1994; Miller, 2013).

Intelligibility and severity are inherently different but together provide a holistic understanding of speech ability. Severity measures assess the accuracy of speech sounds, the presence of phonological processes, and produce outcomes describing the number of incorrect speech sounds. Such severity outcome measures include an objective score that provides a means for comparison to non-disordered speech. Intelligibility measures ask how much of or how well the speech was understood by the listener regardless of correct production. While the evidence does indicate that severity measures and intelligibility measures are strongly correlated; severity measures cannot replace intelligibility measures (Shriberg, Austin, Lewis, McSweeny, &

Wilson, 1997). Measures of severity, such as percent consonants correct (PCC), and standardized assessments for articulation and phonological processes are, however, often used to infer intelligibility based on the number and types of sound errors (Logan, 2010). Although these methods are frequently adopted for diagnosis, speech productions do not have to be accurate to be considered intelligible. Speech can often be produced with errors and still be understandable to the listener.

Children who are particularly impacted by decreased intelligibility are those with phonologically based disorders (Hodson, & Paden, 1983). A child may present with the common phonological process of velar fronting where plosive sounds articulated posteriorly in the oral cavity are replaced with anterior sounds (e.g., initial /k/ is replaced by a /t/). While this would impact the outcome of a severity measure, the predictability of the substitution would result in a minor impact on intelligibility (Dodd 1995). On the contrary, idiosyncratic speech errors, such as initial consonant deletion, may be less predictable, resulting in decreased intelligibility (Bankson, et al., p177). While research supports decreased intelligibility in children with greater presence of phonological processes, little research has explored the possibility of phonological processes contributing to listener confusion or misunderstanding. Certain phonological processes may contribute greatly to unintelligible speech while other processes, although frequently occuring, may not significantly affect intelligibility (Hodson & Paden, 1983 p.63-64).

When evaluating the presence and severity of a phonological disorder, regular practice is to calculate the frequency of occurrence of a phonological processes following the administration of a single-word evaluation (Hodson, 2004). When the percentage-of-occurrence is 40% or greater, it is considered to be an established linguistic pattern not likely to be remediated without therapy if it occurs beyond expected developmental trajectories (Hodson, & Paden, 1983 pp. 69-

74). As children age, the frequency of occurrence of phonological processes should decrease resulting in more intelligible speech. The greater the number of active phonological processes present beyond age expected norms the more severe the SSD. By the age of four, most children will approximate adult-like intelligibility with minimal phonological processing errors (Hodson & Paden, 1981). The correlation between the frequency of errors, presence of phonological processes and unintelligible speech leads clinicians to assign an impressionistic percentage describing the amount of speech that is intelligible during evaluation (Skahan et al., 2007). However, this is approach is not a direct assessment of intelligibility.

Potential Listener Bias in Experienced Listeners

One limitation to clinician-directed intelligibility inferences is the potential biases of an experienced listener. Listeners experienced in understanding a disordered population's speech may find a speech sample of that population to be more intelligible than an inexperienced listener (Flipsen, 1995). This is largely due to the ability for listeners to habituate to the disordered speech (Kent, 1996). Auditory illusion, a phenomenon common in listeners, occurs due to the natural tendency for the auditory system to restore missing or degraded acoustic information in order to comprehend the intended message regardless of the true acoustic presentation (Warren, 1976). Additionally, children's reported intelligibility levels should reflect the functional amount of speech that people in their environment understand (Flipsen, 1995). Intelligible speech is most necessary for experienced and inexperienced listeners, who are likely to have different levels of understanding ability than a clinician knowledgeable in disordered error patterns. Expert listener intelligibility inferences are not the most valid or reliable measurements of functional intelligibility (Kent, 1996).

Towards Explanatory Models of Intelligibility in Speech Disorders

Although improvement of intelligibility for functional communication is a primary goal in remediation of SSD in children, clinical practice has often focused on the treatment of phonological processes as the primary means for improving intelligibility. While the degree of intelligibility and disorder severity measurement, share many of the same factors, few methods have been devised to quantify the effect of different phonological processes on intelligibility of child speech for the purpose of target selection. Prioritizing speech sound intervention is commonly informed by the administration of a standardized phonological assessment protocol such as the Hodson Assessment of Phonological Patterns-3 (HAPP-3). The HA PP-3 was designed to explicitly evaluate the speech of highly unintelligible children. Since, Phonological Deviation Averages (PDAs) have been determined to be significantly correlated with the Percent Consonant Correct- Revised metric (Kwiatkowski & Shriberg, 1992; Shriberg, et al., 1997) and error patterns are weighted according to their negative impact on intelligibility, clinicians rely on such tools in the absence of direct intelligibility measures that are clinically efficient. While a correlation can be drawn between intelligibility and the overall presence of phonological disorders, further investigation into causal factors, that explain how different phonological processes contribute to intelligibility, would provide a more informative approach to determining intervention targets for children.

Hodson explored the relationship between the type of phonological processes present and type of child speaker: unintelligible or intelligible (Hodson & Paden, 1981). Two groups of age-matched child speakers, 3 to 8 years old, were assigned into either the intelligible or unintelligible speaker group determined by parent, SLP, pediatrician, or teacher reports of each child's general communication success. Both groups of child speakers' phonological processes were measured using, The Assessment of Phonological Processes (Hodson & Paden, 1981).

Trained researchers recorded the severity of phonological processes from word transcriptions. It was observed that all child speakers presented with the following phonological processes: cluster reduction, stridency deletion, stopping, liquid deviation and assimilation. Because these processes were reported in both intelligible and unintelligible speakers' speech, they were said to not be large contributors to intelligibility. On the contrary, ten of the most severely unintelligible speakers presented with at least one the following phonological processes at varying degrees: velar deviations, backing, final consonant deletion, syllable reduction, prevocalic voicing, and glottal replacement. Although the level of intelligibility of either speaker group was never directly measured, the significant presence of phonological processes unique to the unintelligible speaker group reflect the predicted relationship between the prominent type of phonological processes and unintelligible speech.

A study performed by Billman (as cited in Hodson and Paden 1991) obtained data from 15 children ages 3;2-6;2 aiming to examine if the presence of some phonological processes had a greater impact on intelligibility than others. The study identified low measures of intelligibility in children with greater production of phonological processes. Individual scores revealed prevocalic singleton omissions and backing were most highly correlated with decreased intelligibility, but liquid /l/ and /r/ errors did not influence intelligibility significantly. Billman also aimed to explain intelligibility deficits by investigating the relationship between intelligibility and the percentage-of-occurrence of phonological process when observing speech patterns of six participants determined to have the least intelligible speech in the study. Stridency deletion and consonant sequence reduction were found to be the phonological processes with the greatest relationship to intelligibility (Billman, 1986. Retrieved from Hodson, & Paden, 1991). By identifying some error types as more prominent in unintelligible speakers than intelligible

speakers these studies lead the authors to question whether all speech production errors make an equal impact on intelligibility. An investigation is needed in order to evidence the effect of specific errors and if some may contribute to intelligibility more than others.

Kent (1989) explored an explanatory approach to assessment of speech intelligibility for adult speakers with dysarthria. Intelligibility was measured across several phonetic contrastive categories to obtain an error profile reflection the most frequently occurring phonetic contrast pairs (Kent, Weismer, Kent, & Rosenbek, 1989). The single word speech stimuli represent nineteen different phonetic contrasts with paired phonemic variations that result in subtle acoustic variation depending on the phonetic contrast category. Intelligibility was measured by the listeners' ability to either recognize the intended word or confuse it with the phonetic contrast pair. In this way, distinguishable phonetic contrasts productions were identified as causing intelligible or unintelligible speech (Kent et al., 1989).

To investigate this explanatory model, phonetic contrast pairs were recorded by twenty-five speakers with Amyotrophic Lateral Sclerosis (ALS). All nineteen phonetic contrasts were tested and analyzed in regard to participant production of the single words. Listeners scored the speaker stimuli via closed-set word recognition. Listener experience was not reported. Intelligibility was analyzed between contrasts groups and results indicated that some error profiles attributed specific phonetic contrasts as contributing to intelligibility more than other acoustic parameters (Kent et al., 1990). Error profiles identified that the stop-nasal and initial glottal null contrasts contributed the most to unintelligible speech. Additionally, variability across subjects with ALS was identified, indicating that the error profiles may be different within a similarly disordered group. These results explain a phenomenon regularly accepted by most clinicians; while speakers of the same disorder may be equally unintelligible they present with

differing errors contributing to intelligibility (Weismer et al., 1992). The fact that two similarly disordered speakers may appear equal when using average intelligibility scores but are actually unique in the type of errors that produce unintelligible speech is a very important clinical finding. It demonstrates that the use of explanatory intelligibility models, rather than quantification of overall intelligibility level, directs clinicians to immediate therapy targets of an individual's most unintelligible errors.

Ansel and Kent further applied the explanatory intelligibility assessment model when examining speech of adult speakers with dysarthria resulting from mixed cerebral palsy (Ansel & Kent, 1992). They combined results from an acoustic analysis of tongue motor movements impacting formant frequencies of dysarthric speech with results from intelligibility testing. This pair of acoustic-motor and perceptual speech studies used phonetic contrastive categories including: syllable initial voicing, syllable-final voicing, stop-nasal consonant, fricative-affricate consonant, high-back vowel, high-low vowel, and tense-lax vowel. The influence of specific contrasts on intelligibility by testing single CVC words was measured by word recognition and interval scale ratings made by eight trained listeners with varying experience levels. Acoustic contrast types were analyzed in order to explore their ability to predict intelligibility. It was found that 62.6% of the intelligibility deficits correlated with fricative-affricate, front-back vowel, high-low vowel, and tense-lax vowel phonetic contrast (Ansel & Kent, 1992). This correlation explains which acoustically measured speaker errors impacted intelligibility the most.

Speakers with dysarthria are characterized by inadequate motor movement of the tongue. When measured independently, the inadequate motor movement resulted in small deviations from healthy speakers. When the acoustic motor movement was paired with an intelligibility measurement it became clear that the intelligibility level of those acoustic-motor differences in

fricative-affricate, front-back vowel, high-low vowel, and tense-lax vowel phonetic contrasts made a large impact on intelligibility (Ansel & Kent, 1992). This study was able to identify the motor movements most significantly impacting intelligibility. The results support the use of perceptual intelligibility measurement with minimally contrasted word pairs as a diagnostic tool.

Klien and Flint (2006) sought to identify the effect of different phonological processes on intelligibility. They equalized error patterns common in child speech through adult speech productions with assigned intensity and type of phonologically disordered speech. Intelligibility was measured by college student listeners via open-set word recognition scores. Results indicated stopping of fricatives and final consonant deletion contributed to intelligibility more than velar fronting when the process occurred at a mild-moderate intensity, 15% to 30% of the time. However, when the frequency of occurrence was severe, 49% to 51%, all three processes, velar fronting, stopping of fricatives, and final consonant deletion, affected intelligibility equally (Klien & Flint, 2006). Phonological processes did not remain constant across levels of severity. While this study identified some differing effects of phonological processes on intelligibility, results of this study are insufficient in describing how different error types contribute to the intelligibility of child speakers. See Table 1 for a summary of explanatory intelligibility studies.

Table 1 Explanatory Intelligibility Studies

Study	Year	Participants	Stimuli	Listener	Method	Contributors to intelligibility
Hodson and Paden	1981	Child Intelligible and Unintelligible	Single words	Trained graduate assistants	Severity: transcribing correct/incorrect at the phoneme level	velar deviations, backing, final consonant deletion, syllable reduction, prevocalic voicing, and glottal replacement
Billman	1986	Child Disordered	unknown	unknown	Intelligibility: unknown	prevocalic singleton omissions backing stridency deletion consonant sequence reduction
Kent	1990	ALS	Phonetic contrasts	unknown-experience	Intelligibility: closed set WRS	stop-nasal, and initial glottal null contrasts
Ansel and Kent	1992	Dysarthric Speakers with Cerebral Palsy	phonetic contrast	eight trained listeners with varying experience levels	Intelligibility: open set WRS, and difficulty rating via IS	fricative-affricate consonant contrasts and the front-back, high-low, and tense-lax vowel contrasts
Klien and Flint	2006	Adults controlled phonological processing errors mimicking child speech	Sentence list	unknown-experience college students	Intelligibility: open set WRS	stopping of fricatives final consonant deletion

Notes. Chronological depiction of explanatory intelligibility studies.

Intelligibility Measurement Regarding Experienced Listeners

Other methods of identifying the effects of intelligibility have been explored using experienced listeners. Hodson and Paden investigated the ability for experienced listeners, such as parents and teachers, to categorize participants into “unintelligible” and “intelligible” study groups. Experienced listeners reported on their ability to understand a child’s speech throughout everyday circumstances (Hodson & Paden, 1981). Kwiatowski and Shriberg (1992) studied the ability of experienced listeners to accurately understand disordered children’s speech. This was assessed by reviewing each caregiver’s ability to accurately gloss their child’s disordered connected speech sample. Caregivers were allotted unlimited viewings of their child’s video and audio tapes while completing open set WRS at the conversation level in order to create glosses of their child’s speech sample. Regardless of assumptions at the time of this study, it was found that

caregivers experienced difficulty in accurately understanding their own child's speech (Kwiatowski & Shriberg, 1992). This study identified the need for an intelligibility assessment capable of describing experienced listener difficulty in understanding disordered, child speech. The Intelligibility in Context Scale (ICS) is a 7-item parent-report measure that assesses the level of intelligibility relating to listeners across varying levels of familiarity to the child (McLeod, Harrison, & McCormick, 2012). Parent ratings on this scale were found to be valid and reliable tools for identifying children in need of a speech sound disorder evaluation (McLeod, 2015). While these assessments of intelligibility have successfully assigned levels of intelligibility to children they are not explanatory measures and lack information describing what errors contribute to intelligibility in child disordered speech.

Measurement of Intelligibility

Direct measurements of intelligibility provide a quantitative approach to determining how well the speech signal is understood (Kent et al., 1994). While direct measures provide greater validity and reliability than the previously discussed severity measures and experienced listener informal estimates, much disagreement remains amongst intelligibility assessment methods (Miller, 2013). Intelligibility can be directly measured using signal independent or signal dependent methods. Signal independent intelligibility measures use information including the acoustic speech signal, visual information, and environmental context including direct observation of the subject. Alternatively, signal-dependent intelligibility measures account for only the acoustic speech signal information. This is collected by listening to the speech sample without accompanying visual or contextual information (Miller, 2013). Signal-dependent measures of intelligibility are reported as being more valid than signal independent measures due to the elimination of uncontrolled human listening principles of phonemic restoration in which

the listener identifies the speaker's intended word regardless of its true production (Kent, 1996). For the purposes of this study, we will focus on signal-dependent intelligibility methods.

Signal Dependent Intelligibility Methods

Quantitative, signal-dependent intelligibility measures assess the speaker's overall degree of intelligibility by either the average of listeners' ability to recognize spoken words or their rating of intelligibility. The frequency of stimuli understood correctly out of the total stimuli tested yields a percentage of words recognized, the Word Recognition Score. The Word Recognition Score (WRS) can be taken at the single word or sentence level where each listener orthographically transcribes the understood message from the speech sample. The speaker's intended production must be known in order to code in a binary fashion whether the listener correctly understood or misunderstood intended message. This can be tested in an open-set format in which the listener transcribes the "real word they think they heard" or closed-set where the listener chooses "what word they think they heard" from a set of choices (Gordon-Brannan, 1994). The frequency of stimuli understood correctly out of the total stimuli tested yields a percentage of word recognized, the Word Recognition Score. Regardless of the accuracy of production, WRSs measure the listener's ability to comprehend the speaker's utterance through the information received via acoustic speech signal.

Signal-dependent intelligibility measures may also employ the use of two different types of rating scales: interval scale (IS) and direct magnitude estimation (DME). Scaling measures are popularly used because of their ability to directly measure the listeners ease in understanding the speaker (Miller, 2013). Interval scale measures are taken by a listeners' response on a 5, 7, or 9-point equally appearing scale used to assess the listener's degree of ease in understanding the speaker. DME is a ratio scale in which the equality of ratios is determined on the degree of

understandability. DME measures are taken via a continuous medium in reference to a learned speech sample by using an assigned point of intelligibility (Schiavetti, Metz, & Sitler, 1981). Unlike the interval scale, DME does not force the listener to choose a number within a linear partition. Rather, DME allows each listener to assign a number that is proportional to their perception of previously heard speech samples. This is done one of two ways: the first is by providing a referent speech sample and intelligibility rating in which all listeners rate speech samples in proportion to the first rating. The second includes the listener rating the first speech sample without an assigned referent but rating all preceding speech samples in correspondence to their first rating (Schiavetti et al., 1981).

When choosing types of measurement ratio is preferable to interval because ratio measurements allow for greater statistical functions (Stevens, 1951). Stevens found that listeners have difficulty dividing their perceptions into equally appearing, linear, intervals. This listener difficulty negatively impacts the validity of interval scales because the lower half of the scale is often subdivided at a greater rate (Stevens & Glanter, 1957). Inadequate use of the entirety of the scale can be explained by the natural mathematical difference between interval and ratio scales. The perceptual nature of intelligibility is similar to that of pitch, brightness, and loudness. These perceptual judgments are not made in regular linear intervals and are best measured by a ratio scale (Stevens, 1986). For these reasons intelligibility is most accurately measured using a ratio scale such as DME rather than an interval scale.

Limitations to Signal-Dependent Intelligibility Measures: WRS & DME

Both WRS and DME possess obstacles that dissuade frequent use as outcome measures in clinical settings. In either the closed or open set, word recognition testing requires the clinician to use a previously selected word or sentence list. Because the SLP must know the words being

tested, WRS is not suitable for spontaneous speech samples that are more representative of everyday speech (Gordon-Brannan, 1994). DME is often not used because it requires listeners to be trained to use a single spoken stimulus and corresponding score as a reference in order for the scale to be used correctly. In addition, DME and WRS measures require three or more inexperienced listeners to score or rate each word tested. Recruiting multiple listeners, preparing the speech samples for listener judgment, and conducting the listening environment require extensive time, compensation, and scheduling demands. WRS and DME, while direct and useful, require more resources than are regularly available in the clinical setting (Gordon-Brannan, 1994, Miller, 2013).

Towards Establishing Clinical Measures of Intelligibility in Children

Intelligibility assessment methods requiring multiple listeners such as WRS and DME have been regularly used in the research setting. However, they still have not been widely adopted in most clinical settings due to the time and resources required for these assessment approaches. One reservation may be due to the requirement of collected and prepared recorded stimulus items. Recordings need to be made in quiet environments using quality recorders. Preparation of recordings may also require manual removal of the clinician's speech requiring knowledge in using acoustic analysis tools and visual interpretation of the speech signal. Advances in technology over time have reduced the cost and resources required for data preparation of digital recordings (Ertmer, 2010).

Another factor contributing to the difficulty of intelligibility assessment is the need for multiple listeners assessment of each speaker's stimuli (Ertmer, 2010). Each listener must listen and judge speech stimuli resulting in a time and resource intensive process. Additionally, many intelligibility studies have employed the use of both experienced and inexperienced listeners to

rate speaker intelligibility bringing to question the ecological validity and consistency of the approaches. Studies with inexperienced listeners have been described as more ecologically valid due to their naiveté to the disordered population's speech patterns (Flipsen, 1995; Kent, 1996; Warren, 1976). Using either approach presents difficulties in recruiting due to the need for multiple listeners, scheduling of the listeners, and compensation. One approach used to address this limitation has been the use of multiple channel amplifiers which allow multiple listeners to listen at the same time through individual headsets (Ertmer, 2010). Most recently researchers have turned to using crowdsourcing as a method for efficient data collection from listeners who are inexperienced representatives of the general population (Byun, 2014; Lansford, Borrie, & Bystricky, 2016; Mayo, Aubanel, & Cooke, 2012; Parson, Braga, Tjalve, & Oh, 2013).

Crowdsourcing

Crowdsourcing is a method of obtaining information through online recruitment of a large number of non-expert listeners. Crowdsourcing is a practical mechanism for streamlining the listener recruitment and speech intelligibility assessment processes due to the compensation and data collection abilities of a web-service such as Amazon Mechanical Turk. Workers (internet users) complete jobs, in this case, intelligibility judgments, in tasks called HITs (Human Intelligence Tasks). Workers receive minor compensation for timely and quality completion of HITs. Use of such a system does not require listeners to leave their own home to complete the study. Additionally, crowdsourcing procedures allow large datasets to be completed simultaneously without any participant scheduling. While some may argue that experimentation over the internet may not allow for adequate controls of extraneous variables, intelligibility testing using crowdsourced listeners results in similar findings to those in a natural environment due to the large and variable listener population. Thus, use of such a large, diverse population of

listeners results in a more ecologically valid measure of intelligibility (Byun, Halpin, & Szeredi, 2015).

Crowdsourcing has been used and validated in multiple studies across health-science fields. A study completed in 2015 investigated the ability of valid intelligibility ratings to be completed via crowdsourcing (Byun et al., 2015). One-hundred words containing misarticulated /r/ sounds were tested by 205 crowdsourced listeners. AMT listeners' intelligibility scores matched "gold standard trained listeners". Findings of this study support crowdsourcing as both a valid and effective form of intelligibility listener judgment (Byun et al., 2015). Lansford, Borrie, and Bystricky (2016) used crowdsourcing to compare listener perceptual training completed at home and online to the known effect of that same training in the laboratory setting. Laboratory collected scores were compared to the second set of listener word recognition scores who were trained and responded via crowdsourcing. All listeners were instructed to orthographically transcribe phrases of real English words they thought they heard. Score comparison between these two groups of listeners resulted in consistent and accurate word recognition scores of AMT workers. This study provides support for use of crowdsourcing as an ecologically valid mechanism for accurate, reliable, and efficient speech intelligibility measurement (Lansford, et al., 2016).

III. Manuscript

Exploring an Explanatory Child Speech Intelligibility Model Using Phonetically Contrasted Word Productions

Introduction

Articulation and phonological disorders are among the most widely treated school-aged disorders by speech language pathologists (SLP). Eight to nine percent of young children have articulation or phonological disorders (NIDCD, 2016). Improvement of speech intelligibility is a primary aim in remediation of articulation or phonological disorders and is described as a fundamental aspect for successful communication (Connolly, 1986). It is widely agreed upon by SLPs that measuring intelligibility is a critical component in an evaluation of speech disorders (Hustad, Oakes, & Allison, 2015; Miller 2013). However, national survey data indicates that as many as 75% of SLPs estimate intelligibility without the use of any standardized protocol (Skahan, Watson, & Lof, 2007). Despite the large influence of intelligibility level on evaluation and treatment of children with speech sound disorders, there is no universal speech intelligibility measure for assessment of child speakers with speech-sound disorders (Hustad, 2018; Miller, 2013).

Speech Intelligibility and Measurement in Children with Speech-Sound Disorders

As a method of classifying the degree of intelligibility, clinicians often assign an impressionistic percentage describing the amount of speech that is intelligible during an evaluation (Logan, 2010). The revised third edition of The Goldman Fristoe Test of Articulation includes an intelligibility measure for connected speech (Goldman & Fristoe, 2015). The

examiner listens to each sentence and rates the speech on four-item rating scale (1) good, (2) fair, (3) poor, or (4) no response. Percentages of intelligible sentences can be compared to same age peers. Such clinician-made judgments are then compared to general intelligibility levels described by age related norms. Subjective clinical ratings of intelligibility are vulnerable to inconsistencies between raters (Ertmer, 2010; Gordon-Brannan, & Hodson, 2000; Kent, Miolo, & Bloedel, 1994; Klein, & Flint, 2006; Miller, 2013). Clinical impressions of intelligibility may also be influenced by experience listening to speech sound disordered speech causing it to be more intelligible than it is for inexperienced listeners (Flipsen, 1995). This experienced listener advantage is due to the ability for listeners to habituate to disordered speech error patterns (Kent, 1996). It is likely that clinicians will understand disordered speech more than inexperienced listeners, thus positively skewing intelligibility measurement. Another disadvantage of impressionistic intelligibility ratings is the failure to explain the aspects of speech production that may cause decreased intelligibility.

More objective analysis of intelligibility has focused on deriving intelligibility from measures of phonological development and consonant production accuracy. The Hodson Assessment of Phonological Patterns and Phonology (HAPP-3) and Percentage of Consonants Correct (PCC) were reported as the most common assessment tools used by school-based clinicians to obtain measures of intelligibility (Hodson, 2004; Logan, 2010; Shriberg, Austin, Lewis, McSweeny, &Wilson, 1997). It should be noted that these are measures of word production that provide information about deviant patterns of speech production. Such measures explain severity and are not direct measures of intelligibility (Logan, 2010). The HAPP-3 is designed to explicitly evaluate the speech of highly unintelligible children who often present with multiple phonological processes (Hodson, 2004). The correlation between highly active

phonological processes measured by the HAPP-3 and decreased intelligibility directs clinicians toward therapy targets of errors with the greatest occurrence. It is important to consider that regardless of the amount or consistency of errors made, multiple errors by the same speaker may not have an equal impact on intelligibility (Billman 1986; Dodd, 1995). The ability to distinguish phonological processes based on their contribution to intelligibility is not measured or known through accuracy assessments. Knowledge of error impact on intelligibility is pertinent for highly unintelligible children with multiple equally-active errors, because all errors cannot be targeted in therapy with the same urgency and priority. While a correlation can be drawn between intelligibility and the overall presence of phonological disorders, further investigation into causal factors, that explain how different phonological processes contribute to intelligibility, would provide a more informative approach to determining intervention targets for children.

Explanatory Intelligibility Assessment

Few methods have been devised to quantify intelligibility of child speech (Flipsen, 1995; Hustad, 2018; Klien & Flint, 2006; Miller, 2013). More systematic methods for measurement of intelligibility have been of interest to researchers for several decades (Flipsen, 1995; Klein & Flint, 2006; Kwiatkowski & Shriberg, 1992; Lousada, Jesus, Hall, & Joffe, 2014; McLeod, Harrison, &McCormack, 2010; Speake, Stackhouse, & Pascoe, 2012). Intelligibility as related to types of speech sound errors has been explored in the adult motor speech population. In adult speakers with dysarthria, explanatory intelligibility studies have found that specific phonetic-based error types have a greater impact on intelligibility than other errors (Kent, Weismer, Kent, & Rosenbeck, 1989). The Diagnostic Intelligibility Test (DIT) is described as a type of explanatory intelligibility assessment of intelligibility (Kent et al., 1989). Diagnostic intelligibility testing seeks to explain the contribution of specific phonetic pairs, categorized by

manner and place of productions, on intelligibility. Intelligibility is measured by the listeners' ability either to recognize the intended word or to confuse it with the phonetic contrast pair. In this way, the assessment aims to link acoustic speech production with intelligibility ratings (Kent et al., 1989). By identifying errors types contributing the most to intelligibility, explanatory intelligibility models direct clinicians to immediate therapy targets of an individual's most unintelligible errors, rather than quantification of overall intelligibility level alone (Ansel & Kent,1992; Kent et al., 1990).

Although evidence indicates phonetic contrasts differ in adult populations with motor speech disorders, implications for child speech sound disorders has yet to be explored using Kent's explanatory intelligibility methods. Investigations finding differences between prominent errors types and speaker intelligibility level have led to the need for explanatory intelligibility models to be applied to children with speech sound disorders (Hodson & Paden, 1991). Studies examining the relationship between phonological processes and intelligibility reported a relationship between prominent phonological processes and intelligibility levels (Hodson & Paden, 1991). Cluster reduction, stridency deletion, stopping, liquid deviation, and assimilation were determined to be equally present in unintelligible and intelligible children. Velar deviations, backing, final consonant deletion, syllable reduction, prevocalic voicing, and glottal replacement were present in severely unintelligible speakers (Hodson & Paden, 1991). Decreased intelligibility has also been attributed to prevocalic singleton omissions and backing while liquid /l/ and /r/ errors did not influence intelligibility significantly (Billman, as cited in Hodson and Paden 1991). Klien and Flint (2006) found stopping of fricatives and final consonant deletion to contributing to intelligibility more than velar fronting when the processes occurred at a mild-moderate intensity, 15% to 30% of the time (Klien & Flint, 2006). While these findings

support the idea that error types and level of severity can influence intelligibility differently, they do not investigate the functional impact of error types thus a need for further research is indicated. Identifying and distinguishing errors that contribute to intelligibility deficits in children with speech sound disorders could guide clinical treatment planning toward targeting specific elements of phonetic contrast categories that influence unintelligible speech in children with speech sound disorders.

The purpose of this study is to 1) employ word sets from the Diagnostic Intelligibility Test (Kent, 1989) to develop an explanatory model of intelligibility reflecting error patterns consistent in children with speech sound disorders, 2) recruits naïve listeners using crowdsourcing to investigate the effects of speaker error type on the general population, 3) measure the impact of phonemic contrast type on listeners' ability to identify the word and their impression of degree of intelligibility. The development of explanatory intelligibility models, while advanced in its functional outcomes, can be hindered by the requirement to recruit multiple inexperienced listeners. Recruiting listeners through a crowdsourcing platform such as Amazon Mechanical Turk (AMT) increases access to inexperienced listeners in the general population, resulting in a more ecologically-valid measurement of intelligibility reflecting impressions of listeners that a child may encounter in real-world environments (Byun, 2014; Byun, Halpin, & Szerdi, 2015; Lansford, Borrie, & Bystricky, 2016; Mayo, Aubanel, & Cooke, 2012; Parson, Braga, Tjalve, & Oh, 2013).

To develop an explanatory model of speech intelligibility in preschool age children, we addressed the following research questions:

(a) Is there a relationship between the accuracy of word production and mean intelligibility scores? It is hypothesized that there is a negative correlation between intelligibility and frequency of errors (Kent, 1992).

(b) Is there a difference in rating of intelligibility by phonetic contrast error type? It is hypothesized that phonetic contrast types of fricative-affricate, stop-affricate, final cluster-final singleton contrast, and stop-fricative will be more unintelligible to the listeners in child disordered speech rather than non-disordered speech (Klien & Flint, 2006; Ansel & Kent, 1992).

(c) Does phonetic contrast type and error frequency predict intelligibility? It is hypothesized that phonetic contrast by type and error frequency predict intelligibility.

Method

Speech Samples

Speech samples were selected from an ongoing study to collect speech samples of children with and without disorders, The Speech Evaluation and Exemplars Database (SEED) (Speights, Boyce, & Willoughby, 2018). The speech samples were recorded by children recruited from a local community early education center. SEED speech samples were recorded under an approved IRB protocol allowing speech samples to be maintained in a public speech database and retrieved for later research use. Speech samples were recorded in a quiet room. Sound levels were measured prior to each recording session to determine if the environmental noise level was below 40 dBA SPL (Williams, Zhou, Stewart, & Knott, 2016). Speech samples were recorded at a 44K sampling rate at 24-bit depth using a handheld H6N recorder with cardioid XLR MOVO LV402 microphones.

Speech samples obtained were word in eight phonetic contrast categories from the DIT including 1) Stop-Fricative, 2) Stop-Affricate, 3) Final Cluster-Final Singleton, 4) Fricative-Affricate, 5) Alveolar-Palatal, 6) Front-Back Vowels, 7) High-Low Vowels, and 8) Initial Cluster-Initial Singleton (Kent, et al., 1989). These contrasts have been associated with decreased intelligibility in children with phonological based disorders (Bankson, et al., 2013; DuHadway & Hustad, 2012; Skahan, Watson, & Lof, 2007).

Nine children with varying levels of speech sound development participated as speakers. Child speakers selected from the database ranged from 3 years 4 months to 5 years 5 months (*M= 4.43*) of age. The sample comprised 5 males and 4 females. Children were evaluated for speech disorders using the Diagnostic Evaluation of Articulation and Phonology (Dodd, Hua, Crosbie, & Ozanne, 2002). Scores are on a scale of 10 and a standard deviation of 3. A score of 7 is one standard deviation below the mean and was used as the criterion for determining with disorder classification. Six children were determined to have non-disordered speech and three to have disordered speech. Phonological processes observed, based on a preliminary analysis of phonetic contrast categories, revealed the presence of stopping of affricates, final consonant deletion, velar fronting, cluster reduction, and backing. All participants included in the study demonstrated: (1) bilateral hearing at 20dB for 0.5kHz, 1kHz, 2kHz, and 4kHz (20dB pass at those four frequencies using Beltone Audio Scout portable audiometer with fitted headphone cups); (2) spoke American English as their primary language; (3) possessed oral communication that included at least one word utterances.

Child speakers were placed into three groups low, mid, and high based on whole word production accuracy related to Proportion of Whole-Word Correctness (PWC) measures. PWC was calculated from transcripts orthographically transcribed by two trained graduate

students. Each student independently completed the transcriptions without prior knowledge of the stimulus list items. PWC is a measure that determines the proportion of words produced correctly out of entire sample set. Binary scoring of the transcriptions coded as “0” for incorrect transcriptions (the transcription not matching the intended word) and “1” for correct transcriptions (the transcription matching the intended word) (Ingram, 2002). A third graduate student broke transcription disagreement. An interrater reliability statistic of .84 was calculated using the word recognition transcription agreement. The agreed upon transcriptions provided each speaker with a proportion of whole word correctness (PWC) score calculated by comparing the number of words produced correctly to the total words produced. PWC percentages were compared to PCC categories, mild > 85%, mild-moderate 65%-85%, moderate-severe 50%-65%, and severe <50% (Shriberg et al., 1997) to inform classification of intelligibility level. Children with a PWC above 85% were considered to be in the high accuracy speaker group (SG) . When PWC fell between 84% and 50%, children were assigned to the mid SG. Those whose PWC fell below 50% were considered to be low SG (See Table 2).

Table 2 Speaker Groups (SG)

	Speaker	**% Whole Word Correct (PWC)**	**Age**	**Sex**	**Disorder**
High accuracy= 100%-85%	H-1	88%	4_4	F	ND
	H-2	87%	4_2	M	ND
	H-3	85%	4_10	M	ND
Mid accuracy= 50%-84%	M-1	83%	4_1	M	ND
	M-2	70%	5_7	M	SSD
	M-3	55%	3_8	F	ND
Low accuracy= 0%-50%	L-1	38%	3_10	F	SSD
	L-2	24%	3_4	M	ND
	L-3	6%	5_5	F	SSD

Notes. Speakers categorized based on percentage of whole words correct.

Preparation of Speech Samples: Materials and Procedure

Sound file sets of the entire stimulus word list were created for each of the speaker groups (High, Mid, and Low). The speaker group sets were then counterbalanced and randomized in order to create 3 sound file stimulus lists later presented to the listeners. Counterbalancing ensured that sound files from every speaker and every word in the stimulus list were distributed evenly across the three lists. At least 10 sound files produced by each of the 9 child speakers were included in each list and thus reduced learning effects of consistent child speech patterns. The sound files in each of the three stimulus lists (Lists 1-3) were then individually randomized to control for order effects. The beginning of each list contained the same 8 sound files of single syllable word productions from the Clinical Assessment of Articulation and Phonology produced by eight of the child speaker participants (Secord & Wayne, 2013). These initial speech samples increased listener practice opportunities and provided consistent stimulus items across different lists. Each list then included its counterbalanced and randomized 92 phonetic contrast sound files differing between each list. Within each of the three lists, no speaker's file was presented more than once, and every sound file was analyzed across seven listeners for intelligibility measurement averages.

Adult Listeners

Prospective listeners from the United States were recruited through the Amazon Mechanical Turk (AMT) crowdsourcing platform. AMT enlists workers (internet users) to complete jobs called HITs (Human Intelligence Tasks). AMT workers were given access to our research experiment link, Intelli-turk©. Workers who selected the Intelli-turk© link and agreed to complete the HIT were assigned a token number and an associated confirmation

code for de-identified administrator task review and compensation. Inclusion criteria required listeners to be at least 19 years of age, non-hearing-impaired, inexperienced with child speech, and speakers of American English as their first language. Following informed consent, individuals self-identified as being inexperienced with child speech by responding "no" to the following questions: Do you have a child who is currently 2-7 years old? Do any of the following apply to you? Pre-school or elementary faculty, a child instructor of any kind, a nanny/caretaker or babysitter, spend more than 10 hours a week listening to children ages 2-7 talk. There were 3 subjects who did not meet requirements and were excluded from the study. One subject was excluded due to failure to meet questionnaire requirements. Two subjects were excluded due to failure to pass the WIPI criteria. 21participants included in the study represented a broad demographic of listeners who did not consistently listen to child speech. Listeners were recruited from 16 US States, included 10 males and 11 females, and ranged in age from 22-72 years old (*M=35.05*). Listeners were compensated monetarily through AMT for their participation.

Listeners were instructed to be in a quiet place and to use headphones or a headset with the volume set at a comfortable listening level. Listeners were required to verify that their computer and headphones were functioning properly before proceeding. Speech recognition ability was screened within the Intelli-turk© web application using the Word Intelligibility Picture Identification (WIPI) Test (Ross & Lerman, 1971). Although this word recognition task was initially designed for children, it has been used to assess listener performance in adults for experimental purposes (Bradley & Sato, 2004; Ishikawa et al., 2017; Lenhardt, Skellet, Wang, & Clark, 1991; Papso & Blood, 1989). Listeners included in the study scored at least 95% on the WIPI.

The listening experiment included 100 words from eight phonetic contrast categories. Listener progress was tracked through the individual de-identified token numbers and confirmation codes generated through the Intelli-turk© administrator platform. Following verification of complete participation, listeners were compensated through AMT.

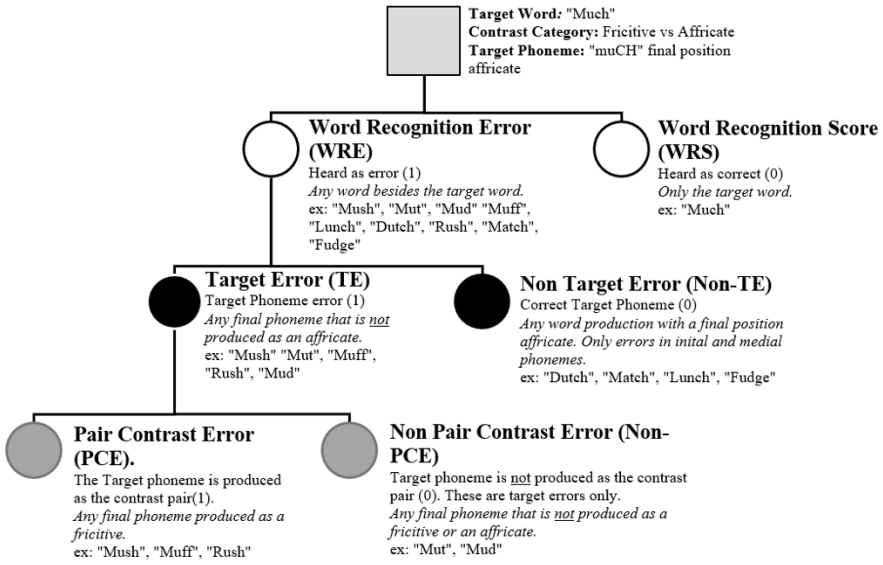
Data Preparation

Listeners' responses were retrieved from a secure SQL server database. Each data set included the typed orthographic transcription and the DME value selected by the listener. Orthographic transcriptions were coded to derive a final Word Recognition Error (WRE) score. DME measures were obtained using the numeric value associated with the scale. Calculations required the average of 7 listeners scores for the three speaker groups across the 8 phonetic contrast categories.

Two trained graduate research team members scored the word recognition responses to identify the number and type of word recognition errors (WRE) made by the group of listeners. Disagreement was broken by a third graduate researcher. WRE scoring involved 3 criteria illustrated by the decision tree in Figure 1. In first criteria, level 1 of the tree, binary coding was used to identify the incidence of unrecognized word productions, WRE. The second criteria of the decision tree indicates if the target phoneme contributed to listener misunderstanding, it was identified as a Target Error (TE), or if the target phoneme was not errored, Non Target Error (Non-TE). The third level of the decision tree, identified the cause of the TE as due to listener confusion with the contrast pair, referred to as phonetic contract errors (PCE), or due to any other misunderstanding, TE. WRE was calculated across the eight phonetic contrast categories for each speaker group. To isolate causes of WRE. The TE total was subtracted from the WRE total,

producing the Non-Target Errors (Non-TE) factor. Next, the PCE total was subtracted from the TE total, to isolate the PCE from other TEs.

Figure 1 Word Recognition Scoring Flow Chart



Note. Methodology of word recognition scoring.

DME values were obtained from the location where the scale marker was placed by the listener for each item. The ability for the phonetic contrast error types and original speaker group accuracy measurement to predict listener difficulty level was measured by comparing binary coded WRS incidence of unrecognized word productions and the proportion of errors to the DME scores. DME scores reflect how the listener rated the intelligibility of each item.

Statistical Analysis

A one-way ANOVA was performed to determine if listeners' DME ratings differed between speaker groups categorized according to PWC range (low, mid, high). The relationship between the proportion of total whole-WRE, target error (TE), and phonetic contrast error (PCE) and mean intelligibility scores by DME was examined using a linear regression model. It was hypothesized that there is a negative correlation between intelligibility measured by DME and the proportion of phonetic contrast errors (Kent 1992). Additionally, it was hypothesized that the

phonetic contrast errors would account for a significant amount of the variance in DME. Linear and logistic regression models were tested to explore WRE and phonetic contrast categories as a predictors of DME.

Results

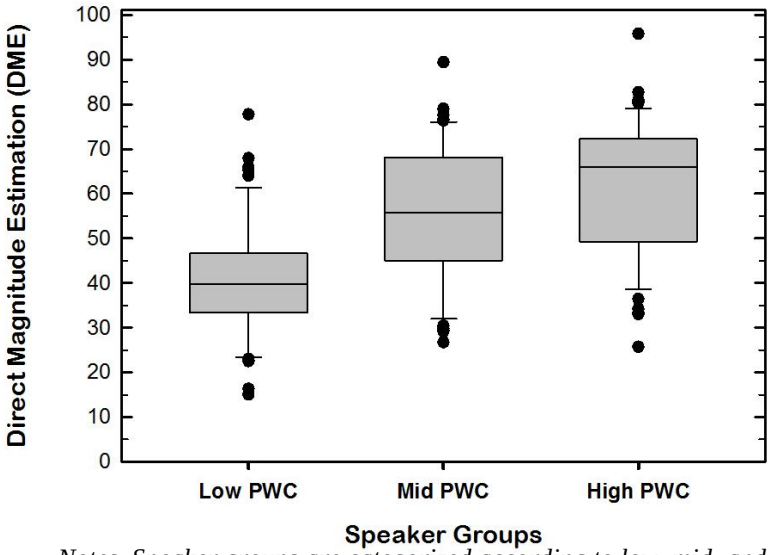
Listeners were trained to perform two tasks (1) Word Recognition (WR) and (2) Intelligibility Rating by Direct Magnitude Estimation (DME) prior to participation in the study. For the WR task, listeners were instructed to first write the word they thought heard, even if it was not a real word, in the text box (Miller, 2013). For the DME task, listeners were introduced to a sliding bar scale in which they rated intelligibility by describing a speech sample from Easy-to-Difficult to understand. DME measures were taken via a continuous medium in reference to a learned speech sample, the referent sound file. Listeners learned the assigned rating point on the scale associated with the reference speech sample. Schiavetti, Metz, & Stiler, 1981). Listeners were trained to listen to the example recording referenced to be at a value of 50 and then mark each subsequent intelligibility rating according to the placement of the learned reference sample (Stevens, 1951). The reference sample was selected and agreed upon by three expert listeners for presenting with average intelligibility within the stimuli set (Schiavetti, et al., 1981). Following the training, listeners completed a 5-item practice module in which they finished the recognition (WR) and rating task (DME).

Out of 24 adults subjects, 3 were eliminated due to failure to meet listener requirements. Twenty-one listeners met all requirements and judged intelligibility by orthographically transcribing the word they thought they heard and rating the level of intelligibility using the Direct Magnitude Estimation scale.

Relationship between Speaker Groups and DME

One-way ANOVA calculation shows a significant difference between the DME scores for speakers in the low, mid, and high accuracy speaker groups (SG) categorized by PWC ([$F(2,12) = 12.76$; $p < .0001$; observed power = .98). Post hoc- Fisher's protect t-LSD multiple comparison tests indicated that listeners reported significantly poorer DME ratings for the low SG (*M*= 41.24; *SE*=1.78; *SD*=13.32) compared to the mid SG (*M*= 55.26; *SE*=2.01; *SD*=15.02) and high SG (*M*=61.88; *SE*=2.01; *SD*=15.02). The linear model reflects a relationship between speaker groups and DME.

Figure 2 Speaker Groups vs. Direct Magnitude Estimation



Notes. Speaker groups are categorized according to low, mid, and high proportion of whole-word correctness as a measure of speech production accuracy. Listeners rated intelligibility according to how intelligible they considered the speech to be to render a DME value. DME rankings were averaged for each group.

Word Recognition Error Types as a Predictor of DME

Multiple regression analysis was performed to investigate the relative contribution of orthographic errors categorized by error type (i.e., Non-TE, TE, PCE) on the intelligibility ratings (DME). Distribution measures and bivariate correlation analyses identified the predictors: PCE, TE, and NonTE. Results of the bivariate correlations among the dependent variable

(DME), and three quantified measures of error types, are provided in Table 2. Bivariate correlation analysis identified a strong inverse relationship among DME and the frequency of errors produced [r =-.90; p < . 001; n= 24]. However the revised predictors derived from the total errors indicated that paired production errors [r = -.53; p = .008; n = 24] and non-TE errors [r = -.50; p = .014 n = 24], both had a moderate inverse relationship to DME. Partial correlations controlling for Non-TEs indicated that TEs were also moderately inversely related to DME scores [r = -.62; p = .002; n=21]. Note 24 represents the average of seven listeners' scores for each of the 24 subcategories in the study (i.e., the data points produced for eight phonetic-contrast categories across three levels of speaker accuracy).

Table 3 Bivariate and Partial Correlations between DME and Proportion of WRE

Variables	*1 DME*	*2 WRE*	*3 Non-TE*	*4 TE*	*5 PCE*
DME	–			*-.62***	*-.55***
WRE	*-.90*	–			
Non-TE	*-.50***	*.52**	–		
TE	*-.37*	.40	*-.30*	–	
PCE	*.53***	.64*	*.10*	*-.10*	–

Note. Values above the diagonal represented partial correlations, controlling for Non-TME.
**Significant at an alpha level of .01 (two-tailed; N = 24).
*Significant at an alpha level of .05 (two-tailed; N = 24).

Altogether, Non-TE, TE and PCE accounted for 83% of the variance in the DME values [ΔR^2 = .81 (adjusted $R2$ = .878); $F(3,20)$ = 34.46, p < .001; observed power = .96]. A hierarchical regression (see Table 4) was conducted to examine this unique contribution and explain variance using the following order of predictors: Non-TE, TE, and PCE. When

controlling for Non-TEs (25%), TEs account for 29% of the variance in DME. Additionally, PCEs accounted for 27% of unique variance in DME of intelligibility.

Table 4 Hierarchical Linear Regression Analysis for DME Value

Variable	*β*	*SE β*	ΔF	*(df1, df2)*	ΔR^{2a}	*Observed Power*
Step 1			*7.12*	*(1, 22)*	*.25**	*.96*
Non-TE	*-31.51*	*11.81*				
Step 2			*13.27*	*(2, 21)*	*.29***	*.99*
Non-TE						
TE	*-42.34*	*9.80*				
Step 3			*27.61*	*(3, 20)*	*.27***	*.99*
Non-TE						
TE						
PCE	*-34.01*	*6.47*				

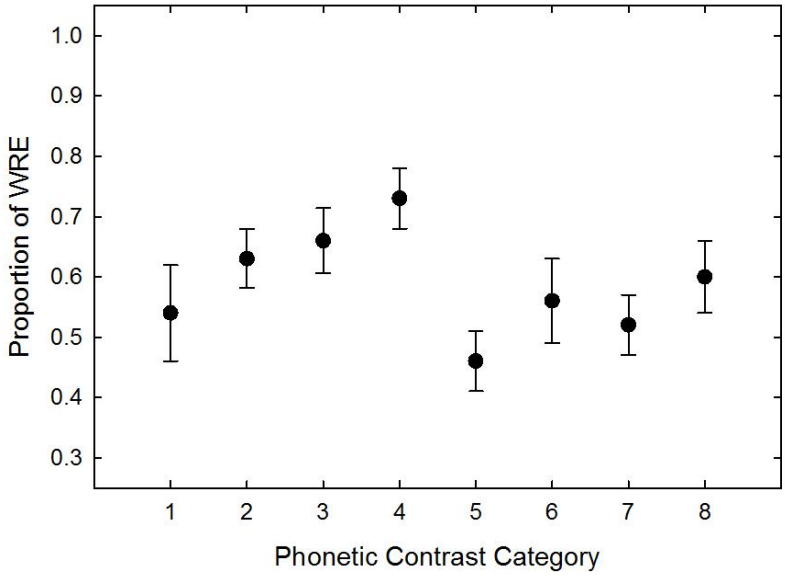
Note.
[a]Total variance accounted for = 81%; Adjusted R^2 =78% (*N* = 24).
**Significant at an alpha level of .001 (two-tailed).
*Significant at an alpha level of .05 (two-tailed).

Proportions of Word Recognition Errors within Phonetic Contrasts Categories

A repeated-measures within-subjects ANOVA on the WRE was completed using NCSS v07.1.21 to determine significant differences in the listeners' orthographic transcriptions for each speaker group across the eight phonetic contrast categories. Categories are referred to by the following numbers: 1) Stop-Fricative, 2) Stop-Affricate, 3) Final Cluster-Final Singleton, 4) Fricative-Affricate, 5) Alveolar-Palatal, 6) Front-Back Vowels, 7) High-Low Vowels, and 8) Initial Cluster-Initial Singleton. The results indicate a significant main effect for SG, [$F(2,12) = 115.94$; $p < 0.001$; $\eta^2 = .95$; observed power = 1.00]. Post hoc- Fisher's protect t-LSD Multiple comparison tests indicated listeners produced significantly larger proportion on WRE for the low SG (M=0.91; SE=0.02; SD=0.12) compare to the mid SG (M=0.45; SE=0.02; SD=0.20) and

high SG (M=0.40 ; SE=0.02; SD=0.15). The results also indicated a main effect for phonetic contrast category [$F(7,42) = 17.81$; $p < 0.001$; $\eta^2 = .75$; observed power = 1.00]. Post hoc analysis, using Fisher's protect t-LSD Multiple comparison tests (*DF*= 42; *MSE* = 0.01; *Critical value* = 2.02) indicated the listeners made significantly more WREs when orthographically transcribing the speech from category 4 (M=0.73) than the remainder of the categories. Category 3, was the second most difficulty category to transcript (M=0.66) with poorer performance than categories 5, 7, 1, 6, and 8. The next category, 2 (M=0.63), produced significantly more WREs than category 5,7,1,6. Category 8 (M=0.60) produced significantly greater proportion of WRE than categories 5, 7, and 1. Categories 6 (M=0.56), 1 (M=0.54), and 7 (M= .52), produced greater proportion of error than category 5. Across all speaker groups, listeners made the least proportion of WREs when orthographically transcribing category 5 (M = 0.46). No other significant differences were identified. See Figure 3 for the visual representation of the mean performance for each phonetic contrast category across the three speaker groups.

Figure 3 Phonetic Contrast Category vs. Proportion of WRE



Notes. The proportion of word recognition errors are measured across each of the eight phonetic categories.

Additionally, a significant interaction was identified between speakers groups and phonetic contrast categories [$F(14, 84) = 5.28$; $p < 0.01$; $\eta^2 = .47$; observed power = .93]. Post-hoc Fisher's protect t-LSD Multiple comparison test indicated that the listeners made significantly more WREs when orthographically transcribing the speech of the low SG for category 1, 2 , 3, 4, 6 and 8 compared to categories 5 (*M* = 0.76) and 7 (M = 0.76; *critical value* = 1.98; *DF*=84; MSE =0.01). For the Mid SG, significantly more WREs were made in categories 4 (*M* =0.79), compared to the remainder of the categories. Additionally, more WREs made in category 3 (M =0.60), compared to categories 1 (M=0.29), 5 (M= 0.36), 6 (M=0.34), 7 (M= 0.34) and 8 (*M* = 0.37) for the mid SG, category 2 (M=0.53) produced significantly more WREs than categories 1, 5, 6, and 7. Categories 1, 5, 6, 7, and 8 did not differ significantly in WREs. For the High SG, a significantly greater proportion of WREs were made in category 2 (*M* = 0.47) than category 5 (*M* = 0.26) and 1 (M=0.34). No other significant differences were identified within each SG. Table 4 provides the means of the proportion of WRE for phonetic contrast categories within each speaker group. Categories listed below the mean in italics indicate critically less WREs were produced for the given speaker groups (column 2, 3 and 4).

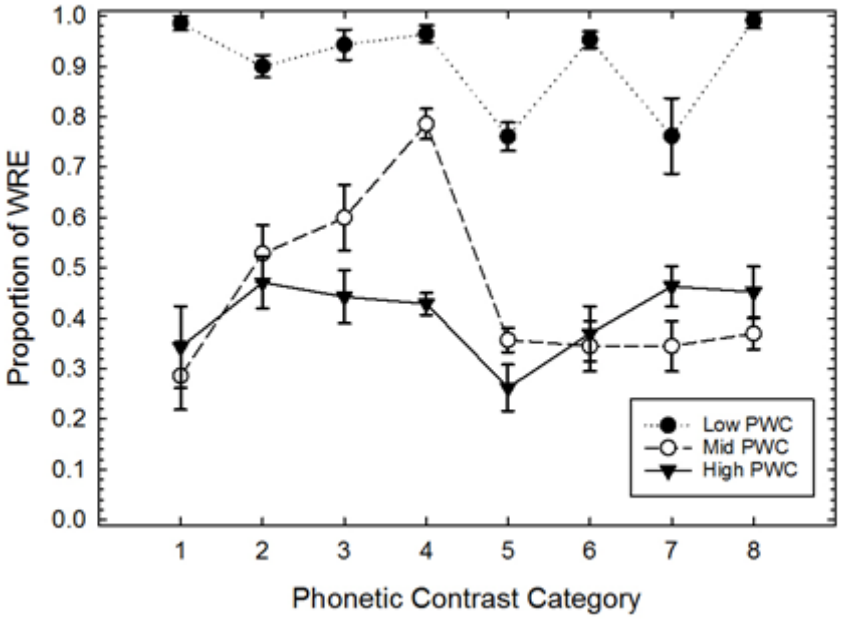
Table 5 Proportion of WRE per Speaker Group

Post-hoc Fisher's protected-t LSD for WRE

	Low	Mid	High
(1) Stop-Fricative			
Mean	**0.99**	0.29	0.34
Critical Diff.	***5,7****		
(2) Stop-Affricate			
Mean	0.90	0.53	**0.47**
Critical Diff.	*5,7**	*1,5,6,7,8**	***5,1****
(3)Final Cluster-Final Singleton			
Mean			
Critical Diff.	0.94	0.60	0.44
	*5,7**	*1,5,6,7,8**	*5**
(4) Fricative-Affricate			
Mean	0.97	**0.79**	0.43
Critical Diff.	*5,7**	***1,2,3,5,6,7,8****	*5**
(5) Alveolar-Palatal			
Mean	0.76	0.36	0.26
Critical Diff.			
(6) Front-back vowels			
Mean	0.95	0.34	0.37
Critical Diff.	*5,7**		
(7) High-low vowels			
Mean	0.76	0.34	0.46
Critical Diff.			*5**
(8) Initial cluster-initial singleton			
Mean	0.98	0.37	0.45
Critical Diff	*5,7**		5*

Note. The mean value is listed for each phonetic contrast category for each speaker group (column 2, 3, 4). Asterisks indicate the other phonetic contrast categories, which had significantly less proportion of PCEs within the PWC speaker group. *Critical value* = 1.99; *DF*=84; *MSE* =0.01

Figure 4 Phonetic Contrast Category vs. Proportion of WRE by SG



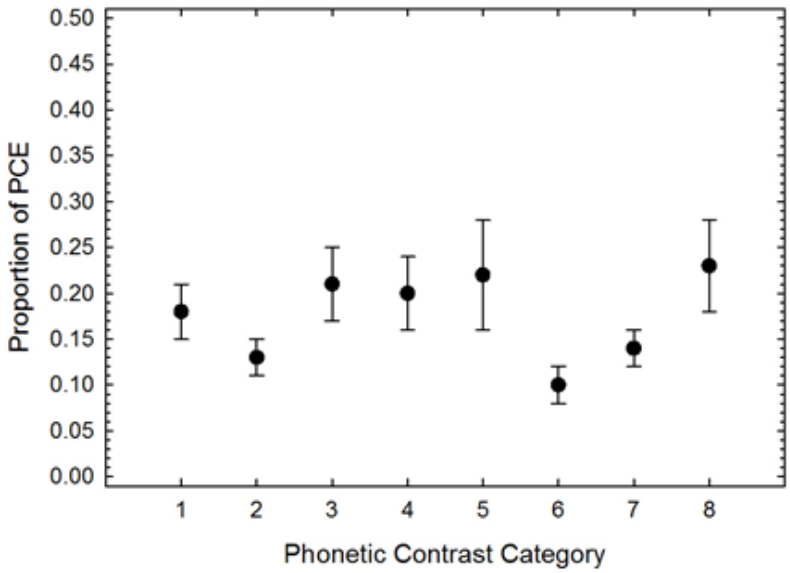
Notes. The proportion of word recognition errors are measured across each of the eight phonetic categories for speakers in the low, mid, and high whole-word accuracy

Proportions of PCE within Phonetic Contrasts Categories

A second repeated-measures within-subjects ANOVA on the PCE was completed to determine significant differences in listeners' orthographic transcriptions for each SG for the eight phonetic contrast categories. The results indicate a significant main effect for SG [$F(2,12) = 68.80$; $p < 0.001$; $\eta^2 = .92$; observed power = 1.00]. Post hoc- Fisher's protect t-LSD Multiple comparison tests indicated listeners produced significantly larger proportion of PCE for the Low SG (M=0.31; SE=0.03; SD=0.20) compared to the Mid SG (M=0.13; SE=0.02; SD=0.14) and High SG (M=0.09; SE=0.01; SD=0.08). Additionally, orthographic transcriptions from the Mid SG produced significantly greater PCEs than the High SG. The results also indicated a main effect for phonetic contrast category, [$F(7,42) = 8.87$; $p < 0.001$; $\eta^2 = .60$; observed power = .99]. Post hoc analysis, using Fisher's protect t-LSD Multiple comparison tests (DF= 42; MSE = 0.01;

Critical value = 2.02) indicated that the listeners made significantly more PCEs when orthographically transcribing the speech from category 8 (M=0.23) than the categories 6, 2, 7 and 1. For categories 5 (M=0.22), 3 (M=0.21), and 4 (M=0.21) listeners transcribed greater PCEs than categories 2, 6, and 7. Next, category 3 produced significantly more PCEs than category 2, 6, 7. Category 1 (M=0.18) produced significantly greater proportion of PCE than categories 6, and 2. Categories 2 (M=0.13), 6 (M=0.10), and 7 (M= .14), produced similar proportion of PCE. No other significant differences were identified. See Figure 5 for the visual representation of the mean performance for each phonetic contrast category across the three speaker groups.

Figure 5 Phonetic Contrast Category vs. Proportion of PCE



Notes. The proportion of phonetic contrast errors are measured across each of the eight phonetic categories.

A significant interaction was identified between speaker groups and phonetic contrast category for PCEs [$F(14, 84) = 32.37$; $p < 0.001$; $\eta^2 = .87$; observed power = 1.00]. Post-hoc Fisher's protect t-LSD Multiple comparison test indicated the listeners made significantly more PCEs when orthographically transcribing the speech of the Low SG for categories 5 ($M = 0.58$)

and 8 (*M* = 0.57) compared to the remainder of the categories (*critical value* = 1.99; *DF*=84; MSE =0.01). Category 3 (M=0.43) had significantly greater proportion of PCEs than categories 1,2,4,6,7. Category 1 (M=0.34) had significantly more PCEs than categories 2,4,6,7. For the Mid SG, significantly more PCEs were made in categories 4 (*M* =0.40), compared to the remainder of the categories. Additionally, more PCEs were made in categories 2 (M =0.10), 3 (M=0.11), 6 (M= 0.14), and 7 (M=0.14) compared to category 5 (M = 0.01). For the High speaker group, listeners produced more PCEs for categories 1 (M=0.13), and 4 (M=0.13) than 6 (M=01), and 8 (M=0.05). Categories 2 (M = 0.11), 3 (M= 0.10) and 7 (M= 0.11) had a greater proportion of PCEs than category 6. No other significant differences were identified within each SG. See table 6 for means and critical differences. See figure 6 following the table for depiction.

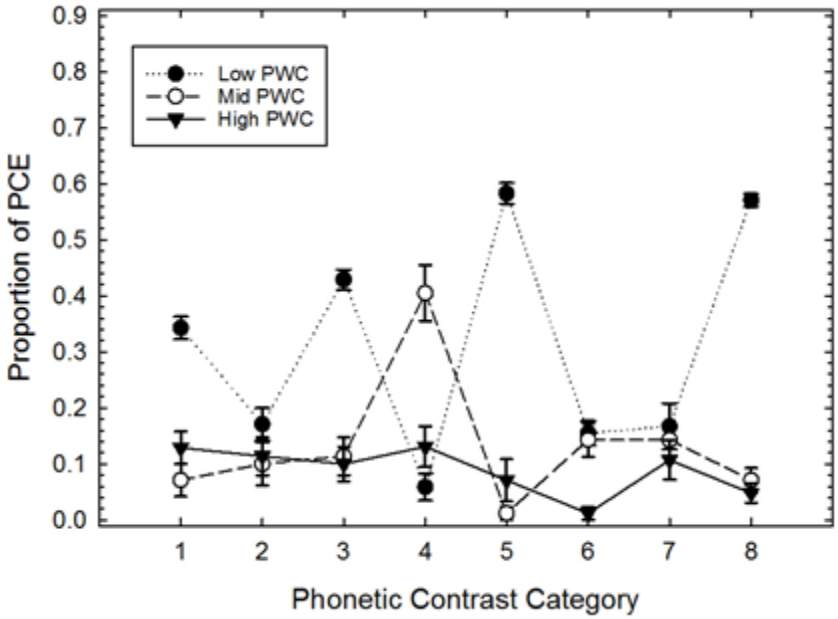
Table 6 Proportion of PCE per Speaker Group

Post-hoc Fisher's protected-t LSD for PCE

	Low	Mid	High
(1) Stop-Fricative			
Mean	0.34	0.07	**0.13**
Critical Diff.	*2,4,6,7**		***6,8****
(2) Stop-Affricate			
Mean	0.17	0.10	0.11
Critical Diff.		*5**	*6**
(3) Final Cluster-Final Singleton			
Mean	0.43	0.11	0.10
)iff.	*1,2,4,6,7**	*5**	*6**
(4) Fricative-Affricate			
Mean	0.06	**0.40**	**0.13**
Critical Diff.		***1,2,3,5,6,7,8****	***6,8****
(5) Alveolar-Palatal			
Mean	**0.58**	0.01	0.07
Critical Diff.	***1,2,3,4,6,7****		
(6) Front-back vowels			
Mean	0.16	0.14	0.01
Critical Diff.		*5**	
(7) High-low vowels			
Mean	0.17	0.14	0.11
Critical Diff.		*5**	*6**
(8) Initial cluster-initial singleton			
Mean	**0.57**	0.07	0.05
Critical Diff	***1,2,3,4,6,7****		

Note. The mean value is listed for each phonetic contrast category for each speaker group (column 2, 3, 4). Asterisks indicate the other phonetic contrast categories, which had significantly less proportion of PCEs within the PWC speaker group. *Critical value* = 1.99; *DF*=84; *MSE* =0.01.

Figure 6 Phonetic Contrast Category vs. Proportion of PCE by SG



Notes. The proportion of phonetic contrast errors are measured across each of the eight phonetic categories for speakers in the low, mid, and high whole-word accuracy groups.

Regression Analysis

Follow up simple liner regressions, were explored to determine if the variance in DME attributed to the three types of WREs Non-TE, TE other than PCE, and PCEs were affected by phonetic contrast category. Within each phonetic-contrast category three simple linear regressions were completed and are displayed in Table 7. Due to the small sample size, hierarchical regressions analysis could not be completed for each category therefore results should be interpreted with caution, as unique variance is not reported below. For category 1, Non-TE accounted for 47% of the variance in the DME values [ΔR^2 = .47 (adjusted $R2$ = .43); $F(1,19)$ = 15.94, p = .001; observed power = .99]. PCE accounted for 35% of the variance in the DME values [ΔR^2 = .35 (adjusted $R2$ = .31); $F(1,19)$ = 10.05, p = .005]. For category 2, TE other than PCE accounted for 22% of the variance in the DME values [ΔR^2 = .22 (adjusted $R2$ = .18); $F(1,19)$ = 5.32, p = .032; observed power = .xx]. For category 4, TE other than PCE accounted

for 21% of the variance in the DME values [ΔR^2 = .21 (adjusted $R2$ = .17); $F(1,19) = 5.04$, p = .037]. For category 5, PCE accounted for 22% of the variance in the DME values [ΔR^2 = .22 (adjusted $R2$ = .18); $F(1,19) = 5.40$, p = .031]. For category 6, Non-TE accounted for 50% of the variance in the DME values [ΔR^2 = .50 (adjusted $R2$ = .18); $F(1,19) = 19.04$, $p < .001$; observed power = .99]. Finally, for category 8, PCE accounted for 53% of the variance in the DME values [ΔR^2 = .53 (adjusted $R2$ = .50); $F(1,19) = 21.14$, $p < .001$; observed power = .99].

Table 7 Summary of Simple Regression Analyses for Variables predicting DME for each category (N= 21).

	Non-TE				TE				PCE			
	B	*SE B*	R^2	*F*	*B*	*SE B*	R^2	*F*	*B*	*SE B*	R^2	*F*
1	**-5.02**	**1.26**	**.47**	**15.94****	0.69	5.62	.01	.01	**-8.40**	**2.65**	**.35**	**10.05****
2	-2.32	3.15	.03	.54	**-4.20**	**1.82**	**.22**	**5.32***	-2.32	4.05	.02	.33
3	-3.56	2.08	.13	2.93	-3.19	3.19	.05	1.01	-1.06	2.03	.01	.27
4	2.75	2.91	.05	.90	**-1.70**	**.77**	**.21**	**5.04***	-.89	1.65	.02	.29
5	1.97	4.97	.01	.16	-0.27	3.97	.00	.01	**-2.57**	**1.11**	**.22**	**5.40***
6	**-3.72**	**0.85**	**.50**	**19.04****	N/A	N/A	N/A	N/A	-6.12	3.68	.08	2.76
7	-2.81	1.36	.18	4.26	N/A	N/A	N/A	N/A	-1.87	3.31	.01	.32
8	-0.73	2.34	.01	0.10	-6.60	5.92	.06	1.24	**-3.59**	**.78**	**.53**	**21.14****

Note. * significant at .05; **signifies significance < .025 (Bonferroni adjustment for secondary analysis)
1) Stop-Fricative, 2) Stop-Affricate, 3) Final Cluster-Final Singleton, 4) Fricative-Affricate, 5) Alveolar-Palatal, 6) Front-Back Vowels, 7) High-Low Vowels, and 8) Initial Cluster-Initial Singleton

Discussion

In this preliminary study, the intelligibility determination by listeners with limited exposure to child speech was investigated to explore the relationship between intellgibilty and phonetic categories reflective of common phonological errors produced by young children. The

study was designed as a step towards applying an explanatory model of intelligibility measurement in speech-sound disordered speech to challenges that listeners may encounter when the context is unknown and when they are inexperienced with the speaker. A perception experiment was conducted to determine if inexperienced listeners could provide insight into the types of errors that cause confusion for listeners in social environments in which children may encounter being misunderstood. Difficulties that listeners encounter in understanding speech are often attributed to errors generating from the talker without considering the contribution of the listener in communication breakdowns (McCormack, McLeod, McAllister, & Harrison, 2010). To investigate the type of difficulties listeners in the general population might encounter, participants were recruited through the crowdsourcing platform Amazon Mechanical Turk (AMT). The primary aim of the study was to explore the effect of phonetically contrasted words on the listeners' abilities to recognize the word and rate the intelligibility of the word. Secondly, we sought to determine if the proportion of phonetic contrast types were distributed uniformly or if error rates differed across the three levels of speaker accuracy and the phonetic contrast categories. Our final aim was to determine if specific phonetic contrast categories can predict listeners' rating of DME. The overarching hypothesis of this study was that listeners inexperienced with child speech could rate the speech of children using Direct Magnitude Estimation and reflect different levels of intelligibility in agreement with previous word production accuracy measures.

Speakers With Lower Proportions of Word Correctness Are Rated as Less Intelligible

The results of this study support a correlation between measures of whole-word accuracy and intelligibility (Ingram & Ingram, 2001; McCabe & Bradley, 1973; Schmitt, Howard, & Schmitt, 1983). A significant difference between the low, mid, and high speaker group's word

accuracy was reflected in the DME scores recorded by inexperienced listeners. This finding supports the relationship between decreased word production accuracy and decreased intelligibility. Our findings, support clinical intuition and measures of severity as the first indicator of functional intelligibility level concerns (Shriberg et al., 1997). However, our findings suggest that specific types of errors may contribute differentially to the confusion of inexperienced listeners.

Word Recognition Error Types Predict Intelligibility

Measuring the effect of word recognition types on intelligibility during a clinical assessment may be a cumbersome and time consuming task given the limited time in an evaluation (Tyler et. al, 2002). Results from the study support the hypothesis that different types of errors can contribute to listeners' intelligibility ratings. Specifically, WRE in the low SG found that the following phonetic contrasts were rated with the least intelligibility rated by listeners 1) Stop-Fricative, 2) Stop-Affricate, 3) Final Cluster-Final Singleton, 4) Fricative-Affricate, 6) Front-Back Vowels, and 8) Initial Cluster-Initial Singleton. In the mid SG the following phonetic contrasts were rated with the least listener intelligibility 3) Final Cluster-Final Singleton, and 4) Fricative-Affricate. In the high SG the 2) Stop-Affricate phonetic contrast was rated with the least listener intelligibility.

Word Recognition Errors by Phonetic Category Type

Beyond intelligibility associated with general WREs it was hypothesized that phonetic contrast types of fricative-affricate, stop-affricate, final cluster-final singleton contrast, and stop-fricative would result in greater PCEs due to their association with phonological processing disorders (Ansel & Kent, 1992; Bankson, et al., 2013; Duhadway & Hustad, 2012; Klien & Flint, 2006; Skahan, Watson, & Lof, 2007). Results of this study suggest that speaker group

level of whole-word accuracy is a significant factor for identifying differences across the phonetic category types. Phonetic contrast errors were examined for each speaker group to determine if listeners were challenged by different categories based on the speaker group. An explanatory component of intelligibility was useful for explaining differences according to phonetic contrast categories in speakers with dysarthria due to ALS and Cerebral Palsy (Kent, 1989, Ansel & Kent, 2002). Because PCEs were significant predictors of DME we explored the differences between phonetic contrast groups by speaker groups. Multiple t-tests were used to compare PC groups to one another through two interactions (1) contrast categories and (2) speaker groups.

When rating speakers with low whole-word accuracy, listeners predominantly experienced difficulty in Stop-Fricative, Final Cluster-Final Singleton, Alveolar-Palatal, and Initial Cluster-Initial Singleton. Speech from the mid SG yielded a different profile of frequent contrast confusions: Fricative-Affricate, Front-Back Vowels, and High-Low Vowels. When listening to speech with few accuracy errors listener intelligibility produced more errors in the categories of: Stop-Fricative, Fricative-Affricate, Front-Back Vowels, and High-Low Vowels. Findings in this study reveal that listeners found different phonetic contrast to be more difficult to understand (or impact intelligibility more) as a result of different levels of speaker whole-word production accuracy. Findings in this study are similar to those in which an explanatory intelligibility study of adult speakers with dysarthria revealed differing phonetic error proportions between speakers within a diagnostic group (Ansel & Kent, 2002). This may be due to intelligibility differences in speakers across different severity levels. Additionally, these findings agree with those describing stopping of fricatives as related to unintelligible child

speakers (Klein & Flint, 2006). However, more exploration is needed in order to compare these findings to those in other similar studies.

Clinical Implications

In clinical practice, there is no standard measure of intelligibility that explains the kinds of difficulties that listeners who are inexperienced with child speech may encounter. Typically, intelligibility is inferred from standardized measures of severity or from subjective ratings of perceptual impressions of intelligibility. Word-recognition batteries can highlight phonetic characteristics of word structure that could prove problematic for listeners. These however may differ according to the degree of intelligibility of the speech signal. The Diagnostic Intelligibility Test (Kent, 1989) uses an explanatory model, and attempts to address specific sounds or sound contrasts that can provide insights into why a listener is having difficulty understanding beyond an overall subjective rating of intelligibility.

We applied an explanatory model by selecting phonetic contrasts that reflect common phonological processes that occurred in child speech and reflect phonological disorders that affect intelligibility. We interpret these results to mean that listeners inexperienced with child speech productions identify differences in intelligibility categorically. Findings were consistent with clinical impressions of level of intelligibility drawn from severity measures of Whole-Word Correctness and rankings according to Percent Consonant Correct.

Children present with multiple equally active phonological processes or multiple articulation errors. Literature points SLPs in many directions of best therapy methods but current phonological process remediation strategy does not include when to pick errors to target if multiple processes are equally active 40% of the time, and may not address functional speech gains in intelligibility. These results are able to contribute to the next steps of speech-sound

remediation and point to the functional implications of error types on inexperienced listeners. For speakers across different accuracy levels certain phonetic contrast categories seemed to contribute more to difficulties encountered by inexperienced listeners and would therefore be higher priority targets for therapy in cases of multiple processes.

Intelligibility measures in this study generally agree with the accuracy measures used to originally classify the speakers. However, in few categories the mid SG was found to be more intelligible than the high SG. This discrepancy draws significance to the need for a more regularly used direct intelligibility measure in clinical settings. In these instances, the accuracy scores alone were not able to completely convey the functional impact of the speech sound disorder on intelligibility.

Limitations and Future Directions

This study reports a preliminary investigation towards establishing an explanatory model of intelligibility for preschool age children. This study included 9 speakers and 21 listeners. Although speakers were categorized according to whole-word production accuracy, we did not control for age and disorder types. Increasing the sample size of speakers would allow for examining the effects of age and disorder type. Specifically the small sample size may have impacted overall intelligibility testing in which measurements in the mid accuracy group surpassed or were increasingly similar to those in the high speaker group. These findings warrant future consideration of classification of children within the mid-range of whole-word accuracy. Retrospective review of the Whole Word Correctness scores revealed the two out of the three speakers in the mid SG performed closer to the high accuracy level rather than the low accuracy group. A larger sample size of speakers with whole-word correctness scores that fall closer to the median of the mid SG may yield phonetic contrasts that differ between the mid and high speaker

groups. Additionally use of a single word for intelligibility measurement may not be representative of the amount of speech required for a listener to understand a conversation. While this limits the ability to generalize findings beyond single word intelligibility scores, use of a single word was necessary in this study in order to evaluate WR and DME.

Further study should include repeated measures with speech pathology clinicians as listeners to investigate differences between the inexperienced and experienced perception of speaker intelligibility. Further study would also benefit from additional listeners to further explore the influences of phonetic categories on intelligibility ratings. Variability between listeners was not a focus of this study, it was however observed that measures of direct magnitude estimation varied across listeners. We intend to explore machine learning to address the variability in DME ratings as means for understanding human imprecision in the rating of intelligibility.

Acknowledgements

The research was supported by funding from the College of Liberal Arts and the Department of Communication Disorders at Auburn University. This project was supported by multiple undergraduate and graduate students in the Technologies for Speech-Language Research Lab. Bayley Smith and Bailey Evans provided transcriptions of words for calculation of whole-word accuracy. Caroline Willis aided in the preparation and organization of speech samples for the study. Alexandra Brooks and Emily Hanner assisted in scoring listener responses. We greatly appreciate their support.

References

ASHA practice portal: Speech sound disorders-articulation and phonology. (n.d.). Retrieved from https://www.asha.org/PRPSpecificTopic.aspx?folderid=858993532 1&section=Assessment.

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research*, *35*(2), 296–308.

Billman, K. S. (1986). *Phonological processes and intelligibility of spontaneous utterances in young children* (Unpublished doctoral dissertation) Available from *Targeting intelligible speech: A phonological approach to remediation*, Hodson and Paden 1991.

Bankson, N. W., Bernthal, J. E., & Flipsen, P. (2013). Speech sound assessment procedures. *Articulation and phonological disorders: Speech sound disorders in children*, 177-211.

Byun, T. M., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*, *53*, 70–83.

Connolly, J. H. (1986). Intelligibility: a linguistic view. *International Journal of Language & Communication Disorders*, *21*(3), 371–376. https://doi.org/10.3109/13682828609019848

DuHadway, C. M., & Hustad, K. C. (2012). Contributors to intelligibility in preschool-aged children with cerebral palsy. *Journal of Medical Speech-Language Pathology*, *20*(4). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299463/

Dodd, B., (1995). *The differential diagnosis and treatment of children with speech disorder.* San Diego, CA: Singular Publishing Group, Inc.

Dodd, B., Hua, Z., Crosbie, S., & Ozanne, A. (2002). *Diagnostic Evaluation of Articulation and Phonology: DEAP*. London, England: Psychological Corporation Ltd.

Edition, F., & Bauman-Waengler, J. (2012). *Articulatory and phonological impairments: A clinical focus*. Oxnard, CA: Pearson.

Ertmer, D. J. (2010). Relationships between speech intelligibility and word articulation scores in children with hearing loss. *Journal of Speech, Language, and Hearing Research*, *53*(5), 1075–1086. https://doi.org/10.1044/1092-4388(2010/09-0250)

Flipsen Jr, P. (1995). Speaker-listener familiarity: Parents as judges of delayed speech intelligibility. *Journal of Communication Disorders*, *28*(1), 3-19.

Gelman & Hill (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

Goldman, R., & Fristoe, M. (2015). *Goldman Fristoe Test of Articulation. 3.* Minneapolis MN: Pearson Education, Inc., & PsychCorp (Firm).

Gordon-Brannan, M. (1994). Assessing intelligibility: Children's expressive phonologies. *Topics in Language Disorders*, *14*(2), 17–25.

Gordon-Brannan, M., & Hodson, B. W. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology*, *9*(2), 141–150.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23.

Hodson, B. W., & Paden, E. P. (1983). *Targeting intelligible speech: A phonological approach to remediation*. Austin, TX: College-Hill Press.

Hodson, B. W. (2004). *HAPP-3: Hodson Assessment of Phonological Patterns*. *3.* Austin, TX:

Pro-Ed.

Hustad, K. (2018, November). *1568: Motor Speech Disorders in Children & Adults: Lessons from Development and Degeneration.* Seminar presented at the annual convention of the American Speech-Language-Hearing Association, Boston, MA.

Hustad, K. C., Oakes, A., & Allison, K. (2015). Variability and diagnostic accuracy of speech intelligibility scores in children. *Journal of Speech, Language, and Hearing Research, 58*(6), 1695–1707. https://doi.org/10.1044/2015_JSLHR-S-14-0365

Hustad, K. C., Schueler, B., Schultz, L., & DuHadway, C. (2012). Intelligibility of 4-year-old children with and without cerebral palsy. *Journal of Speech, Language, and Hearing Research: JSLHR*, *55*(4), 1177–1189. https://doi.org/10.1044/1092-4388(2011/11-0083)

Ingram, D., & Ingram, K. D. (2001). A whole-word approach to phonological analysis and intervention. *Language, speech, and hearing services in schools*.

Kent, R.D. (1992). *Intelligibility in speech disorders* Philadelphia, PA: John Benjamins Publishing Company.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, *5*(3), 7–23.

Kent, R. D., Kent, J. F., Weismer, G., Sufit, R. L., Rosenbek, J. C., Martin, R. E., & Brooks, B. R. (1990). Impairment of speech intelligibility in men with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Disorders*, *55*(4), 721–728.

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology*, *3*(2), 81–95.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility

testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499.

Klein, E. S., & Flint, C. B. (2006). Measurement of intelligibility in disordered speech. *Language, Speech, and Hearing Services in Schools*, *37*(3), 191-199.

Kwiatkowski, J., & Shriberg, L. D. (1992). Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss. *Journal of Speech, Language, and Hearing Research*, *35*(5), 1095-1104.

Lansford, K. L., Borrie, S. A., & Bystricky, L. (2016). Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. *American Journal of Speech-Language Pathology*, *25*(2), 233-239.

Logan, N. R. (2010). *Methods used to assess intelligibility in children with phonological disorders: Results of a national survey*. Retrieved from the University of Central Missouri.

Lousada, M., Jesus, L. M. T., Hall, A., & Joffe, V. (2014). Intelligibility as a clinical outcome measure following intervention with children with phonologically based speech–sound disorders. *International Journal of Language & Communication Disorders*, *49*(5), 584–601. https://doi.org/10.1111/1460-6984.12095

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. *Thirteenth Annual Conference of the International Speech Communication Association*.

McCabe, R. B., & Bradley, D. P. (1973). Pre-and Postarticulation Therapy Assessment. *Language, Speech, and Hearing Services in Schools*, *4*(1), 13-22.

McLeod, S., Harrison, L. J., & McCormack, J. (2012). The intelligibility in context scale: Validity and reliability of a subjective rating measure. *Journal of Speech, Language, and Hearing Research*, *55*(2), 648-656.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, *48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

National Institute on Deafness and Other Communication Disorders (2016). *Quick Statistics about voice, speech, language, and swallowing.* Retrieved from: https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language

Parson, J., Braga, D., Tjalve, M., & Oh, J. (2013). Evaluating voice quality and speech synthesis using crowdsourcing. In *International Conference on Text, Speech and Dialogue*. 233-240.

Ross, M., & Lerman, J. (1970). Word Intelligibility by Picture Identification. *Journal of Speech and Hearing Research, 13,* 44-53.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech and Hearing Research*, *24*(3), 441–445.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons.

Secord, W., Donohue, J. A. S. (2002). CAAP: Clinical Assessment of Articulation and Phonology. Greenville, S.C: Super Duper Publications.

Schmitt, L. S., Howard, B. H., & Schmitt, J. F. (1983). Conversational speech sampling in the assessment of articulation proficiency. *Language, Speech, and Hearing Services in Schools*, *14*(4), 210-214.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research, 40*(4), 708–722.

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment

practices for children with suspected speech sound disorders: Results of a national survey. *American Journal of Speech-Language Pathology*, *16*(3), 246–259.

Speake, J., Stackhouse, J., & Pascoe, M. (2012). Vowel targeted intervention for children with persisting speech difficulties: Impact on intelligibility. *Child Language Teaching and Therapy*, *28*(3), 277-295.

Speights Atkins, M., Boyce, S. E., & Willoughby, K. E., (2018). SEED- Speech Exemplars and Evaluation Database. Auburn University Technologies for Speech-Language Research Lab. Permanent URL: http://hdl.handle.net/11200/49140

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). Oxford, England: Wiley.

Stevens, S. S. (1986). *Psychophysics: Introduction to its perceptual, neural and social prospects*. New Brunswick, NJ: Transaction Publishers.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of experimental psychology*, *54*(6), 377.

Tyler, A. A., Tolbert, L. C., Miccio, A. W., Hoffman, P. R., Norris, J. A., Hodson, B., ... & Bleile, K. (2002). Five views of the elephant: Perspectives on the assessment of articulation and phonology in preschoolers. *American Journal of Speech-Language Pathology*, *11*(3), 213-214.

Weismer, G., Martin, R., & Kent, R. D. (1992). Acoustic and perceptual approaches to the study of intelligibility. *Intelligibility in Speech Disorders*, 67–118.

Williams, W., Zhou, D., Stewart, G., & Knott, P. (2016) The practicality of using a smart phone 'App' as an SLM and person noise exposure meter. *Proceedings of ACOUSTICS* Brisbane, Australia.

Yorkston, K. M., Beuklemon, D. R., & Traynor, C. (1984). *Computerized assessment of intelligibility of dysarthric speech.* Austin, TX: Pro-Ed.